Journal of Information Engineering and Applications
ISSN 2224-5782 (print) ISSN 2225-0506 (online)
Vol.7, No.6, 2017
www.iiste.org

Big Data Processing with Apache Spark in Tertiary Institutions: Spark Streaming

Emmanuel Boachie[1] Prof Chunlin Li[2]
Wuhan University of Technology,
School of Computer Science and Technology, Wuhan 430063, China
[1]Kumasi Technical University, Kumasi, Ghana

The work was supported by the National Natural Science Foundation (NSF) under grant (No.61472294, No.61672397)

Abstract
In tertiary institutions, different set of information are derived from the various department and other functional sections. Individual departments and other functional sections in the institutions manage their data separately. This situation has resulted in huge number of different set of data across the various departments in tertiary institutions. There is no centralized data centre where data/information can be retrieved for the management committee when the need arises. In academic institution data captured is restricted to the institution which collected it but centralisation of the various data in the various functional sections does not exist. This makes it difficult for the management committee to take decisions based on relevant information needed. In order to address this problem, we proposed Spark Streaming. Spark Streaming is an element which facilitates processing of live flows of data. Spark streaming will able to capture data in real time, process it and make it available to the management committee when the need arises

Keywords: Spark, Streaming, Big data, Processing, Tertiary, Institution

1. Introduction

Data are increasingly growing in tertiary institutions in Ghana because of rising number of students and courses. It is unproductive or nearly unlikely to use Relational Databases Management Systems to manage this diversity of data streaming into the system. Relational Databases Management system can easily manage homogenous data but not all types of data are homogenous. There are other heterogeneous data in the academic institutions which includes video, audio, pictures, location data, simulations Magnetic Resonance Imaging (MRI) and so on [1]

These data type to be stored and managed are very huge. This means that, it will occupy huge space ranging from the size of petabyte (10^{15}) to yottabyte (10^{24}). The utmost accessible space for SQL 2014 for Business Intelligence, Enterprise, Standard and Web applications is 524 petabytes ($524x10^{15}$), this indicates that, with the exponential growth of data retrieved in the tertiary sector on daily basis, it will be complex to store data with this type of RDBMS.

Information is important to the management committee of every academic institution. With some basic features of processed data being Accessibility (data needs to be simply available to the authorized users in that they can get it in the appropriate structure and at the appropriate moment to meet their wants.), Timely (Timely data is provided the time it is required) and Verifiability (It shows that one can verify it to be certain that it is accurate, maybe by verifying a lot of sources for the similar information.) The procedure of accessing information must also be timely.

In tertiary institutions, different set of information are derived from the various department and other functional sections. There is no centralized data centre where data/information can be retrieved for the management committee when the need arises. In academic institution data captured is restricted to the institution which collected it but centralisation of the various data in the various functional sections does not exist. This makes it difficult for the management committee to take decisions based on relevant information needed. All the departments and other functional sectors manage their data separately and provide them to the management committee when they are asked for. These data are normally provided manually, most at times in hard copies, by the various functional sectors to the management committee. This situation normally prevents the management committee to have relevant information when the need arises for decision making.

In order to address this problem, Spark Streaming should be used. Spark Streaming is an element that facilitates processing of live streams of data. Spark streaming offers an API for controlling data streams that intimately correspond with the spark core`s RDD API, making it simple for developers to study the task and move among applications that control information kept in memory, or coming in real time. Beneath its API, spark streaming has been devised to offer the equal level of error acceptance, throughput and scalability as spark core. Spark streaming can capture data in real time, process them and make it available to the management committee when the need arises.

2. Related Work

Big Data is a knowledge system that is already changing the objects of knowledge and social theory in many fields while also having the potential to transform management decision-making theory [2]. Big Data incorporates the emergent research field of learning analytics, which is already a growing area in education. However, research in learning analytics has largely been limited to examining indicators of individual student and class performance. Big Data brings new opportunities and challenges for institutions of higher education. Long and Siemen [3] indicated that Big Data presents the most dramatic framework in efficiently utilising the vast array of data and ultimately shaping the future of higher education. The application of Big Data in higher education was also echoed by [4], who noted that technological developments have certainly served as catalysts for the move towards the growth of analytics in higher education.

In the context of tertiary education, Big Data connotes the interpretation of a wide range of administrative and operational data gathered processes aimed at assessing institutional performance and progress in order to predict future performance and identify potential issues related to academic programming, research, teaching and learning. Others indicated that to meet the demands of improved productivity, higher education has to bring the tool of analytics into the system. As an emerging field within education, a number of scholars have contended that Big Data framework is well positioned to address some of the key challenges currently facing higher education [5]. At this early stage much of the work on analytics within higher education is coming from interdisciplinary research, spanning the fields of Educational Technology, Statistics, Mathematics, Computer Science and Information Science. A core element of the current work on analytics in education is centred on data mining.

Big Data in tertiary education also covers database systems that store large quantities of longitudinal data on students' right down to very specific transactions and activities on learning and teaching. When students interact with learning technologies, they leave behind data trails that can reveal their sentiments, social connections, intentions and goals. Researchers can use such data to examine patterns of student performance over time from one semester to another or from 1 year to another. On a higher level, it could be argued that the added value of Big Data is the ability to identify useful data and turn it into usable information by identifying patterns and deviations from patterns. [6] indicate that Big Data is now well positioned to start addressing some of the key challenges currently facing higher education. An [36] report suggested that it may be the foundation on which higher education can reinvent both its business model and bring together the evidence to help make decisions about educational outcomes.

From an organisational learning perspective, it is well understood that institutional effectiveness and adaptation to change relies on the analysis of appropriate data and that today's technologies enable institutions to gain insights from data with previously unachievable levels of sophistication, speed and accuracy [7]. As technologies continue to penetrate all facets of higher education, valuable information is being generated by students, computer applications and systems [8].

Furthermore, Big Data Analytics could be applied to examine student entry on a course assessment, discussion board entries, blog entries or wiki activity, which could generate thousands of transactions per student per course. These data would be collected in real or near real time as it is transacted and then analysed to suggest courses of action. As [9] indicated that "[learning] analytics are a foundational tool for informed change in education" and provide evidence on which to form understanding and make informed (rather than instinctive) decisions.

a) The Value of Big Data in Tertiary Education

Big Data can also address the challenges associated with finding information at the right time when data are dispersed across several unlinked different data systems in institutions. By identifying ways of aggregating data across systems, Big Data can help improve decision-making capability [10]

b) Implementation Setbacks

The number of predictable setbacks related with the execution of analytic procedures for Big data in tertiary education are many. A number of these include setbacks related with securing consumers to acknowledge Big data as a channel for applying fresh processes and alter administration. Again, there is an incredible charge related with gathering, keeping and introducing mathematical formulas to extract data, a process which is predisposed to be time wasting and technical. More so, a lot of organizational information schemes are not interoperable, so cumulating managerial information, classroom and online data can present extra setbacks [11].

Spectacular progress in information gathering, processing power, information communication and keeping capacities are facilitating many institutions to combine their variety of databases into data banks. In the era of plentiful information, tertiary institutions related to institutions, healthcare or government sectors have some of the similar grounds for applying analytics, particularly in the sections of financial effectiveness, enlarging local and worldwide effect, deal with innovative monetary support models throughout a altering cost-effective climate, and retorting to the requirements for better responsibility.

Within tertiary institution, information are on the rise, though the majority of it is spread out transversely

desktops, departments and come in a variety of layouts, making it complex to generate or merge. To successfully access this information, the capability to examine varied information sets is required, in spite of their source, and combination data kept in silos within institutions, administering and administrating the information while saving insightful data across databases, is a core demand for application of big data in higher schooling.

Analytics also has the ability to assist students and teachers to be aware of risk signals before menace to studying success achieved. Nevertheless, broad institutional reception of analytics demands an apparent organizational plan and the utilization of analytics software packages [12]

Moreover, information storage implements normal data layouts. Every section will generate outcomes which are standardized with all the other sections, resulting in additional precise information demonstration. Finally, a information storehouse can keep huge quantities of past information that can be willingly for testing and to examine diverse periods of time and patterns in order to make prospect forecasting.

3. Method

a) The settings and Participants

Information needed for this research were gathered from the various departments and other functional sections in Kumasi Technical University-Ghana. The rational being that they are the key instruments that can provide appropriate and relevant data for analysis upon which an Apache Stream framework and processing model can be developed for data processing.

b) Data Collection

The staff from various the departments and other functional sections such as institutional, research administration and curriculum section who are in charge of data processing were the right people to provide relevant data for this research work. They are the staff who are working on the data that come to their end and process them manually and make them available to the management committee when the need arises.

c) Data Analysis

Data that were gathered from the various departments and other functional sections were considered for analysis. Qualitative approach was used for the analysis. The result enabled the researcher to study the true situation on the ground in terms of data mobilization and processing. Based on that Spark streaming architecture has been developed to replace the traditional approach of managing the institutional data for management committee to take decision.

d) Quality Control Procedure

In order to ensure quality work as far as the implementation of Apache Spark; Spark Streaming is concern, the personnel in the Information and Communication Directorate ICT were involved to manage the system by gathering information from the various departments and other functional sections and process them for storage. The ICT Directorate has been placed in the position to make information available to the management committee but not the individual functional sections in the institution by processing the data using apache streaming technology.

4. Results

The findings of the research indicated that the various functional sections and the departments provide information on a hard copy to the management for decision making. There is no centralized data or data bank which contains all the information from the various functional sections in the institution. ICT Directorate too do not have any connection with the functional sections for data mobilization and processing. Therefore, the result of the research undertaken has confirmed the usefulness of the application of Apache Streaming technology in data mobilization and processing which is able to capture data in real time for analytics.

5. Apache Spark Streaming Tool

Apache Spark is a cluster computing platform developed to be swift and universal- idea. On the pace side, Spark widens the well-liked MapReduce model to professionally sustain more sorts of computations, including interactive queries and stream processing. Rapidity is imperative in processing huge datasets, as it indicates the variation involving searching information interactively and waiting minutes or hours. One of the major elements of Spark provides for rapidity is the capability to operate computations in memory, but the system is also well-organized than MapReduce for multifaceted applications operating on disk [13].

The considered method that will be used is Spark outflow which is an element of Spark technology for data processing. Spark SQL is a package for operating with formatted data. It permits querying figures through SQL and the Apache Hive alternative of SQL termed as Hive Query Language (HQL) and it wires numerous origins of information, such as Hive tables, Parquet, and JSON [9]. Outside offering a SQL interface to Spark, Spark SQL permits builders to blend SQL queries with the programmatic information handling backed by RDDs in Python, Java, and Scala, which are a sole appliance, which is joining SQL with multifaceted analytics. This rigid mixing couple with the affluent computing atmosphere offered by Spark enables Spark SQL not like any other

unwrap origin of data storehouse apparatus. Spark SQL has been contained Spark in edition 1.0. Shark happened to be older SQL-on-Spark scheme out of the California University, Berkeley, which customized Apache Hive to operate on Spark. Spark SQL has taken over now to provide advanced incorporation with the Spark engine and language APIs [14].

Spark Streaming is a Spark an element which induces processing of live streams of data. Data streams consist of log files retrieved by creating web servers, or queues of communications covering status current information which clients of a web service normally sent. Spark Streaming offers an API for handling information flows which intimately in line with the Spark key's RDD API, rendering it simple for the developers to study the scheme and progress among applications that maneuver information saved in memory, on disk, or arriving genuine occasion. Beneath its API, Spark outflow was planned to offer similar level of error acceptance, throughput, and scalability as Spark key [13]. Spark comes with in-build machine learning library and Spark streaming which is capable to capture information in genuine occasion for analytics.

5.1. Application

In advanced stage, all Spark application comprises a driver program which initiates a diversity of corresponding functions on a cluster. The driver program consist of one`s application's core functionality and classifies allocated datasets on the cluster, and uses functions to them. In the foregoing for instance, the driver program was the Spark shell itself, and we could immediately enter in the functions we want to operate. Driver programs contact Spark via SparkContext entity that signifies a link to a computing cluster. In the case, a SparkContext is involuntarily formed for you as the variable called sc. probing the sc variable>>>sc<pyspark.context. SparkContext entity at 0x1025b8f90>If we have a SparkContext, we can apply it to develop RDDs. For instance, wsc. toin a file. We can then operate diverse functions on these lines like count (). To operate these functions, driver programs normally administer an amount of nodes termed as executors. If we operate the count() function on a cluster, unrelated devices may calculate lines in diverse series of the file. Since we now operate the Spark shell in the vicinity, it implemented any of its function on a sole device but we can tie the similar shell to a cluster to examine information in corresponding executes on a cluster [15]. The diagram below shows the driver application program of distributed execution in spark.

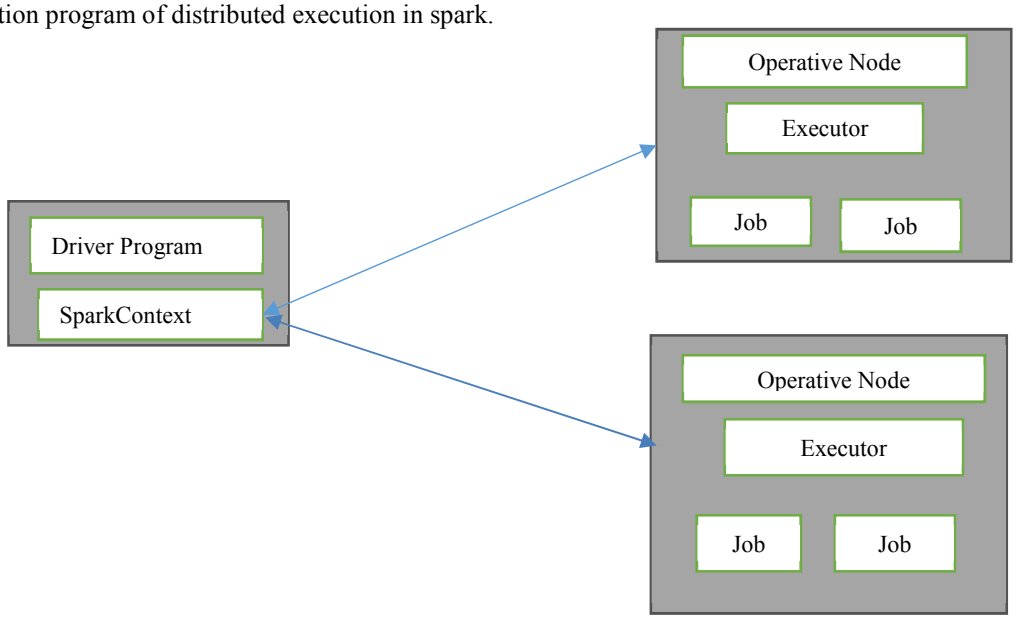

Fig 1. Distributed execution in Spark

5.2 Data Processing Implementation

Spark Streaming executes incremental stream processing applying a model termed as "discretized streams" To execute streaming over spark; we divide the input data into small batches that we frequently merge with state hoarded within RDDs to offer fresh outcomes. Operating streaming computations in this mode has a lot of advantages over traditional distributed streaming with batch and interactive queries.

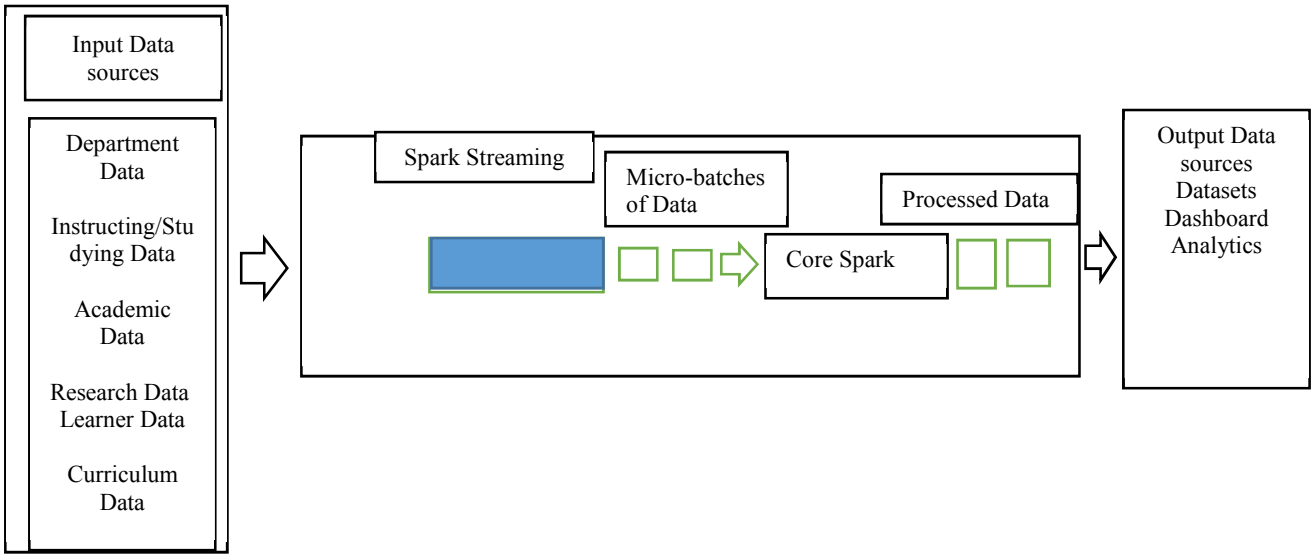

Figure 2. Data Processing

The figure 2. Indicates the processing trend of big data with Spark Streaming technology.

The data that flood the institution on daily basis can now be processed in a real time. Here, a databank will be created after all the data gathered from the various departments and other functional sections in the institutions are processed. This has enabled the management committee to have access to information they need every time for decision making. The implementation of spark streaming technology has addressed the problem of the individual departments and other functional sectors managing their own data and making it available to the management committee.

6. Discussions and Implications

6.1. Evaluation

Big data processing using Apache Spark Streaming system can be evaluated under several criteria. Execution, Storage efficiency, generating effectiveness, and what it offers to the user are some of the criteria used for evaluation. The relevance of these factors is by the system developer and the equally most suitable spark streaming processing procedure and data structure that will be designed for implementation are dependent on the decisions made by the developer.

The efficiency in its Execution is measured by the time it takes a module of the system or the system at large to perform a successful computation. It is measured using C supported systems. Execution efficiency is and always will be a major point of concern for big data processing systems.

6.2. Evaluation Environment

There is peer to peer kind of system architecture that have proved to be a better alternative to the tiresome and less flexible decentralized system. The Spark streaming system has adopted this type decentralized of network structure. There is not really a mother computer or machine that oversees others hence the named peer. All the computers in the spark streaming domain are peers. This makes the transfer of information at any given time possible. Another advantage is that when some machines (computers) somewhere the domain malfunction, the others will work irrespective to that. If this eventuality was to occur in a centralized kind of system, and the victim computer happens to be the main frame, the whole system will shut down and crumble. The peer to peer type of organization structure is easy to maintain and faster in operation compared to the main frame.

6.3 Data sources

Information that interests us as the users of any data that comes from all the departments and other functional sections in the Kumasi Technical University. They include text documents, sensor data, photographic pictures, biological sources, and web pages. Accessing this does requires an intermediary which is common in the process of big data. The data to be processed is then carefully indexed for the later to be successful. This all process can be summarized as data mining in the data management system where information can be made available to the management committee for decision making.

6.4 Parameter Setting

Parameter setting is a mode of attempting to generally and quickly clarify the language acquisition. It is simply a mode for processing huge data with spark streaming to improve continues functionality of the establishment.

This conception offers accounts on why and how we are able to build grammatically correct language in diverse cases without memorizing them in our heads. This is a typical case in the huge data processing with spark streaming.

A parameter set can be fitted to a different environmental setting via evolution simulation and the outcome is still hoped for by the consumer. To sum it up, parameter location makes it likely for the big data processing with spark streaming system to be an institutional based as the intranet which are the fastest route of communication is an institutional based web. It makes it likely for information sharing to bypass language barriers and other unlikely social standards. The most important part of parameter setting theory is to understand how the kind of our language provides enormous short-cuts. Compared to its effect in the big data processing with spark streaming, the consumer doesn't really have to know the approach he or she is applying to obtain the information they are seeking for. That part has been encapsulated. The consumer only types in his or her needs and a short while he gets them anywhere he or she is.

6.5 Metrics

Metrics is the calculation of performance of a given procedure. The metrics in big data are classified into categories. Online metrics calculate the original real-time users' interactions with the exploration system. The offline metrics calculate the significance of the exploration engine via imposing expert moderators to calculate the probability of each outcome to suite the information needs of the consumer. This metrics is estimated applying the set theory knowledge. It applies connections, cardinality symmetric distinction, summation etc. Online metrics is retrieved from data gathered from search logs. The Online metrics such as Click-through rate, session abandonment rate, session success rate, and zero outcome rate. The offline metrics is retrieved from vital ruling incidents. Here, judges are allowed to evaluate by score the effectiveness of each search result. Their score is in binary form meaning either superior or inferior no mediator or they apply a multilevel type of scale of satisfying unlimited search needs of the consumer.

The Normalized discounted collective gain is a metric that calculates the functionality of a suggestion system depending on the ranked significance of entities suggested. It ranks from 0.0 to 1.0. 1.0 stands for the grading of entities.

Parameters: k is the highest figure of entities advocated $GCD_k = \sum_{i=1}^{k} \frac{2^{reli}-1}{\log(i+1)}$ $IGCD_k$ is the highest possible (ideal) GCD for a given que of manuscripts

$$nGCD_k = \frac{GCD\ k}{IGCD\ k}$$

Conclusion

In tertiary institutions, different set of information are derived from the various department and other functional sectors. There is no centralized data centre where data/information can be retrieved for the management committee for decision making when the need arises There is no mechanism in place to analyze accumulated data from all the functional sections in the institution to enable the management committee to get data that are relevant for effective decision making. In order to address this problem, we proposed Spark Streaming which is a component of Apache Spark Technology which is designed to process datasets from various sectors. Spark Streaming is an element that facilitates processing of live streams of data. Spark streaming offers an API for controlling data streams that intimately correspond with the spark core`s RDD API, making it simple for developers to study the project and move between applications that control data kept in memory, or coming in real time. Beneath its API, spark streaming has been devised to offer the same level of error acceptance, throughput and scalability as spark core. Spark streaming is an effective technology that can capture data in real time for processing in tertiary institution to enable the management committee to have access to relevant information to make decision at the right time.

Acknowledgment

The work was supported by the National Natural Science Foundation (NSF) under grant (No.61472294, No.61672397), Key Laboratory of Spatial Data Mining & Information Sharing of Ministry of Education, Fuzhou university (No. 2016LSDMISO5) program for the High-end Talents of Hubei Province Any opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of the above agencies.

References

[1] Okur. C, Buyukkececi, M. (2014). Big Data challeneges in Information Engineering Curriculum . *IEEE*.

[2] Long. P, Siemen, G. (2014) Penetrating the fog: analytics in learning and education. EDUCAUSE Review, 46, 5, 30–40.

[3] Hrabowski, F. A. III, Suess, J. (2010). Reclaiming the lead: higher education's future and implications for technology. EDUCAUSE Review, 45, 6 (November/December 2010). Retrieved October 30, 2014, from

http://www.educause.edu/library/ERM1068
[4] Hrabowski, F. A. III, Suess, J, Fritz, J. (2011a). Analytics in institutional transformation. EDUCAUSEREVIEW, https://net.educause.edu/ir/library/pdf/ERM1150.pdf
[5] Siemens. G. (2011). How data and analytics can improve education, July 2011. Retrieved August 8, October 30,
[6] OECD (2013). OECD Report: The State of Higher Education 2013. Retrieved March 24, 2014, from http://www.oecd.org/edu/imhe/thestateofhighereducation2013.htm
[7] Rowley, J. (1998). Creating a learning organisation in higher education. Industrial and Commercial Training, 30, 1, 16–19.
[8] Borgman, C, Abelson. H, Dirks. L, Johnson. R, Koedinger. K, Linn. et al (2008). Fostering learning in the networked world: The cyber learning opportunity and challenge, a 21st century agenda for the national science foundation. Arlington, VA: National Science Foundation http://www.nsf.gov/pubs/2008/nsf08204/nsf08204.pdf
[9] Mayer, M. (2009). Innovation at Google: The physics of data [PARC forum]. Retrieved 11 August 2009, from http://www.slideshare.net/PARCInc/innovation-at-google-the-physics-of-data
[10] Baker, R. S. J. D. & Inventado, P. S. (in press). (2013) Educational data mining and learning analytics.
[11] Daniel, B. K, Butson. R. (2013). Technology enhanced analytics (TEA) in higher education, Proceedings of the International Conference on Educational Technologies, 29 Novemebr–1 December, 2013, Kuala Lumpur, Malaysia, pp. 89–96.
[12] White. C. (2011) Using Big Data for Smarter Decision Making. BI research.
[13] "Spark Release 2.0.0". MLlib in R: SparkR MLlib APIs [..] Python: PySpark MLlib algorithms"
[14] Zaharia. M, Chowdhury. M. F, Michael. J, Shenker. S, Stoica, I. (2014). Spark: Cluster Computing with Working Sets (PDF). USENIX Workshop on Hot Topics in Cloud Computing (HotCloud)
[15] Zaharia. M, Chowdhury. M, Das. T, Dave. A, Ma. J, McCauley. M, Michael J, Shenker. S, Stoica. I. (2014) Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing (PDF). USENIX Symp. Networked Systems Design and Implementation.